

# CURRICULUM VITAE

SVETLANA A. SHABALINA

National Center for Biotechnology Information  
National Institutes of Health  
8600 Rockville Pike, Bldg.38A, Rm.6S604  
Bethesda, MD 20894  
Phone: (301) 594-5693  
E-mail: shabalin@ncbi.nlm.nih.gov

## EDUCATION

**Ph. D. degree in Computational and Molecular Biology** from the Institute of Molecular Biology, USSR Academy of Sciences, Moscow. Thesis: "Textual and statistical analysis of regulatory regions in DNA and RNA"

**M. Sc. degree in Genetics** from the Moscow State University. Thesis: "The computer analysis of *Drosophila* heat-shock genes".

## POSITIONS HELD

2004 - now    Staff Scientist, National Center for Biotechnology Information,  
National Institutes of Health  
1999 - 2004   Research Fellow, National Center for Biotechnology Information,  
National Institutes of Health  
1997 - 1999   Postdoctoral Associate, Section of Genetics and Development,  
Cornell University, Ithaca, NY  
1995 - 1997   Postdoctoral Associate, Section of Ecology and Systematics, Cornell  
University, Ithaca, NY  
1990 - 1995   Group Leader, Institute of Mathematical Problems in Biology, Russian  
Academy of Sciences, Pushchino, Russia  
1987 - 1992   Research Scientist, the same institute  
1985 - 1987   Associate Research Scientist, the same institute

## RESEARCH INTERESTS

Genome organization and evolution of mammalian noncoding genome sequences. Transcriptional and posttranscriptional regulation of gene expression. DNA and RNA signals involved in the regulation of expression levels and tissue-specific expression. Evolution of miRNA genes. RNA-RNA interactions in the regulation of translation and RNA processing. Mutation accumulation and genomic deleterious mutation rates.

## SELECTED PUBLICATIONS

1. Zhang F, Kang Y, Saha S, Shabalina SA, Kashina A (2010). Differential Arginylation of Actin Isoforms is Regulated by Coding-Sequence-Dependent Degradation. *Science*, (in press).
2. Shabalina SA, Ogurtsov AY, Spiridonov AN, Novichkov PS, Spiridonov NA, Koonin EV. (2010). Distinct patterns of expression and evolution of intronless and intron-containing mammalian genes. *Mol. Biol. Evol.*, **27**, 1745-1749.
3. Matveeva OV, Kang Y, Nechipurenko YD, Nemtsov VA, Shabalina SA (2010). Efficient method for the design of siRNAs and shRNAs. *PLoS One*, **5**, e10180.
4. Losada L, Ronning CM, DeShazer D, Woods D, Fedorova N, Kim HS, Shabalina SA, Pearson TR, Brinkac L, Tan P, Nandi T, Crabtree J, Badger J, Beckstrom-Sternberg S, Saqib M, Schutzer SE, Keim P, Nierman WC. (2010). Continuing evolution of *Burkholderia mallei* through genome reduction and large-scale rearrangements. *Genome Biol. Evol.*, **2**, 102-116.
5. Fozo EM, Makarova KS, Shabalina SA, Yutin N, Koonin EV, Storz G. (2010). Abundance of type I toxin-antitoxin systems in bacteria: searches for new candidates and discovery of novel families. *Nucleic Acids Res.*, **38**, 3743-3759.
6. Resch AM, Ogurtsov AY, Rogozin IB, Shabalina SA, Koonin EV. (2009). Evolution of alternative and constitutive regions of mammalian 5'UTRs. *BMC Genomics*, **10**, 162.
7. Nackley AG, Shabalina SA, Lambert J, Conrad M, Gibson D, Spiridonov A, Satterfield S., Diatchenko L. (2009). Low Enzymatic Activity Haplotypes of the Human Catechol-O-Methyltransferase Gene: Enrichment for Marker SNPs. *PLoS ONE*, **4**, e5237.
8. Shabalina SA, Zaykin D, Ogurtsov AY, Gris P, Gauthier J, Shibata K, Sama S, Chivileva I, Belfer I, Spiridonov NA, Max MB, Goldman D, Fillingim RB, Maixner W, Diatchenko L. (2009). Expansion of the Human  $\mu$ -Opioid Receptor Gene Architecture: Novel Functional Variants. *Hum. Mol. Genet.*, **18**, 1037-1051.
9. Ogurtsov AY, Mariño-Ramírez L, Johnson GR, Landsman D, Shabalina SA, Spiridonov NA. (2008). Expression patterns of protein kinases correlate with gene architecture and evolutionary rates. *PLoS ONE*, **3**, e3599.
10. Shabalina SA and Koonin EV (2008). Evolution and origin of the eukaryotic microRNA system. *Trends Ecol. Evol.* **23**, 578-587.
11. Wang P, Lyman RF, Shabalina SA, Mackay TF, Anholt RR. (2007). Association of polymorphisms in odorant-binding protein genes with variation in olfactory response to benzaldehyde in *Drosophila*. *Genetics*, **177**, 1655-1665.

12. Diatchenko L, Nackley AG, Chivileva I, Shabalina SA, Maixner W. (2007). Genetic Architecture of Human Pain Perception. *Trends Genet.*, **23**, 605-613.
13. Resch AM, Carmel L, Marino-Ramirez L, Ogurtsov AY, Shabalina SA, Rogozin IB, Koonin EV. (2007). Evidence of positive and negative selection in synonymous sites of mammalian genes. *Mol. Biol. Evol.*, **24**, 1821-1831.
14. Matveeva O, Nechipurenko Y, Rossi L, Moore B, Saetrom P, Ogurtsov AY, Atkins JF, Shabalina SA. (2007). Comparative analysis of experimental databases and algorithms for functional siRNA design. *Nucleic Acids Res.*, **35**, e63.
15. Nackley AG, Shabalina SA, Tchivileva I, Satterfield KS, Korchinsky O, Maixner W, Diatchenko L. (2006). Common human catechol-O-methyltransferase haplotype modulates protein expression by altering mRNA secondary structure. *Science*, **314**, 1930-1933.
16. Diatchenko L, Anderson AD, Slade GD, Fillingim RB, Shabalina SA, Bhalang K, Sigurdsson A., Higgins T, Sama S, Belfer I, Goldman D., Max MB, Weir B, Maixner W. (2006). Three major haplotypes of the adrenergic receptor  $\beta 2$  define psychological profile and risk of development of a common musculoskeletal pain disorder. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, **141**, 449-462.
17. Shabalina SA, Ogurtsov AY, Spiridonov NA. (2006). A Periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic Acids Res.*, **34**, 2428-2437.
18. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV. (2006). A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biology Direct.*, **1**, 7.
19. Ogurtsov AY, Shabalina SA, Kondrashov AS, Roytberg MA. (2006). Analysis of internal loops within the RNA secondary structure in an almost quadratic time. *Bioinformatics*, **22**, 1317-1324.
20. Shabalina SA, Spiridonov AN, Ogurtsov AY. (2006). Computational models with thermodynamic and composition features improve siRNA design. *BMC Bioinformatics*, **7**, 65.
21. Yampolsky L, Allen CD, Shabalina SA, Kondrashov AS. (2005). Persistence time of loss-of-function mutations at non-essential loci affecting eye color in *Drosophila melanogaster*. *Genetics*, **171**, 2133-2138.
22. Diatchenko L, Slade GD, Nackley AG, Bhalang K, Sigurdsson A, Belfer I, Goldman D, Xu K, Shabalina SA, Shagin D, Max MB, Makarov SS, Maixner W. (2005).

Genetic basis for individual variations in pain perception and the development of a chronic pain condition. *Hum. Mol. Genet.*, **14**, 135-143.

23. Shabalina SA, Ogurtsov AY, Rogozin IB, Koonin EV, Lipman DJ. (2004). Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals. *Nucleic Acids Res.*, **32**, 1774-1782.

24. Shabalina SA, Spiridonov NA. (2004). The mammalian transcriptome and the function of non-coding DNA sequences. *Genome Biol.*, **5**, 105.

25. Matveeva OV, Foley BT, Nemtsov VA, Gesteland RF, Matsufuji S, Atkins JF, Ogurtsov AY, Shabalina SA. (2004). Identification of regions in multiple sequence alignments thermodynamically suitable for targeting by consensus oligonucleotides: application to HIV genome. *BMC Bioinformatics*, **5**, 44.

26. Kolker E, Makarova KS, Shabalina SA, Picone AF, Purvine S, Holzman T, Cherny T, Armbruster D, Munson RS, Kolesov G, Frishman D, Galperin MY. (2004) Identification and functional analysis of 'hypothetical' genes expressed in *Haemophilus influenzae*. *Nucleic Acids Res.*, **32**, 2353-2361.

27. Glazko GV, Koonin EV, Rogozin IB, Shabalina SA. (2003). A significant fraction of conserved noncoding DNA in human and mouse consists of predicted matrix attachment regions. *Trends Genet.*, **19**, 119-124.

28. Silva JC, Shabalina SA, Harris JL, Spouge AS, Kondrashov AS. (2003). Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of Mir and L2-derived sequences within the mouse and human genomes. *Genetical Research*, **82**, 1-18.

29. Matveeva OV, Shabalina SA, Nemtsov VA, Tsodikov AD, Gesteland RF, Atkins JF. (2003) Thermodynamic calculations and statistical correlations for oligo-probes design. *Nucleic Acids Res.*, **31**, 4211-4217.

30. Matveeva OV, Mathews DH, Tsodikov AD, Shabalina SA, Gesteland RF, Atkins JF, Freier SM. (2003). Thermodynamic criteria for high hit rate antisense oligonucleotide design. *Nucleic Acids Res.*, **31**, 4989-4994.

31. Shabalina SA, Ogurtsov AY, Lipman DJ, Kondrashov SA. (2003). Patterns in interspecies similarity correlate with nucleotide composition in mammalian 3'UTRs. *Nucleic Acids Res.*, **31**, 5433-5439

32. Webb CT, Shabalina SA, Ogurtsov AY, Kondrashov AS. (2002). Analysis of similarity within 142 pairs of orthologous intergenic regions of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Nucleic Acids Res.*, **30**, 1233-1239.

33. Kondrashov AS, Shabalina SA. (2002). Classification of common conserved sequences in mammalian intergenic regions. *Hum. Mol. Genet.*, **11**, 669-674.
34. Shabalina SA. (2002). Regions of intermolecular complementarity in *Escherichia coli* 16S rRNA, mRNA, and tRNA molecules. *Mol. Biol.*, **36**, 359-364.
35. Ogurtsov AY, Roytberg MA, Shabalina SA, Kondrashov AS. (2002). OWEN: aligning long collinear regions of genomes. *Bioinformatics*, **18**, 1703-1704.
36. Roytberg MA, Ogurtsov AY, Shabalina SA, Kondrashov AS. (2002). A hierarchical approach to aligning collinear regions of genomes. *Bioinformatics*, **18**, 1673-1680.
37. Shabalina SA, Ogurtsov AY, Kondrashov VA, Kondrashov AS. (2001). Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet.*, **17**, 373-376.
38. Matveeva OV, Tsodikov AD, Giddings M, Freier SM, Wyatt JR, Spiridonov AN, Shabalina SA, Gesteland RF, Atkins JF. (2000). Identification of sequence motifs in oligonucleotides whose presence is correlated with antisense activity. *Nucleic Acids Res.*, **28**, 2862-2865.
39. Yampolsky LY, Webb CT, Shabalina SA, Kondrashov AS (1999). Rapid accumulation of a vertically transmitted parasite triggered by relaxation of natural selection among hosts. *Evolutionary Ecology Research*, **1**, 581-589.
40. Shabalina SA, Kondrashov AS. (1999). Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes. *Genetical Research*, **74**, 23-30.
41. Kondrashov AS, Yampolsky LY, Shabalina SA. (1998). On the sympatric origin of species by means of natural selection. In: *Endless Forms: Species and Speciation* (D. J. Howarth and S. H. Berlocher, eds.), 90-98, Oxford Univ. Press.
42. Shabalina SA, Yampolsky LY, Kondrashov AS. (1997). Decline of fitness in panmictic populations of *Drosophila melanogaster* maintained under relaxed natural selection. *Proc. Natl. Acad. Sci. USA*, **94**, 13034-13039.
43. Nazipova NN, Shabalina SA, Ogurtsov AY, Kondrashov AS, Roytberg MA, Buryakov GV, Vernoslov SE. (1995). SAMSON, a program package for the analysis of biopolymer primary structures. *CABIOS*, **11**, 423-426.
44. Nechipurenko YD, Popov NV, Isaev MA, Shabalina SA, Matveeva OV. (1995). Multiple contact model describing interaction of mRNA with rRNA sites during translation processes. *Biofizika*, **40**, 1208-1213 (in Russian).

45. Matveeva OV, Shabalina SA. (1993). Intermolecular mRNA-rRNA hybridization and the distribution of potential interaction regions in murine 18S rRNA. *Nucleic Acids Res.*, **21**, 1007-1011.
46. Ogurtsov AY, Elkin YE, Shabalina SA. (1993). Calculation of contributions of individual monomeric units into biopolymer functioning. *J. Theor. Biol.*, **164**, 395-401.
47. Shabalina SA, Yuryeva OV, Spiridonov NA, Kondrashov AS. (1988). Comparative analysis of textually similar sequences of DNA regulatory regions. NCBI USSR Acad. Sci., Pushchino, 58 p. (in Russian).
48. Vernoslov SE, Kondrashov AS, Roytberg MA, Shabalina SA, Yuryeva OV, Nazipova NN. (1989). The program package "SAMSON" for the analysis of primary structure of biopolymers. NCBI USSR Acad. Sci., Pushchino, Part 1 - 96 p., Part 2 - 124 p. (in Russian).
49. Vernoslov SE, Kondrashov AS, Roytberg MA, Shabalina SA, Yuryeva OV, Nazipova NN. (1990). The program package "SAMSON" for the analysis of primary structure of biopolymers. *Mol. Biol.*, **v. 24**, 524-529 (in Russian).
50. Shabalina SA, Yuryeva OV, Kondrashov AS. (1991). On the frequencies of nucleotide substitutions in conservative regulatory DNA sequences. *J. Theor. Biol.*, **149**, 43-51.
51. Spiridonov NA, Arkhipov VV, Narimanov AA, Shabalina SA, Shvirst EM, Zverkova LA, Kondrashova MN. (1991). Influence of *Galleria mellonella* larvae preparation and honeybee products on cell cultures. *Comp. Biochem. Physiol.*, **102C**, 205-208.

## **PATENTS**

1. Gesteland RF, Atkins JF, Matveeva OV, Shabalina SA Method, articles, and compositions for identifying oligonucleotides. (November 15, 2004). No. PCT/US2004/038092.

## **REVIEWING OF MANUSCRIPTS**

Genome Research, Cell, Genome Biology, Nucleic Acids Research, PLoS Biology, Trends in Genetics, PLoS One, PLoS Computational Biology, Bioinformatics, BMC Journals.

## TEACHING EXPERIENCE

In 1994 and 1995 I taught the following courses at the Pushchino University and Moscow State University (Russia):

- 1) Introductory Biology for students majoring in mathematics or computer science
- 2) Computer Analysis of Biopolymer Primary Structures.

Before this, in 1989-1992, I participated as a volunteer in teaching an advanced course in Biology (similar to AP) for High School students.

## GRANTS AND AWARDS

- |           |   |
|-----------|---|
| 1993      | Individual grant from the International Science Foundation  |
| 1994      | George Soros fellowship in Biological Sciences  |
| 1991-1994 | Comparative analysis of regulatory regions of human genes, Russian Human Genome Research Foundation                 |
| 1992-1993 | Comparative analysis of 5'-regulatory regions of murine and human genes, Russian State Program of Genetical Studies |
| 1993-1995 | Comparison of homologous tRNAs and rRNAs in related species, Russian State Program in Biodiversity                  |
| 1993-1995 | Selective constraint in non-coding DNAs, Russian Foundation for Fundamental Studies                                 |

## CURRENT AND RECENT RESEARCH PROJECTS

### **Genome architecture of alternative transcription and alternative splicing in mammals**

The diversity of RNA transcripts produced from the same gene locus results in the increased ratio of alternative to constitutive nucleotides in many genes of mammalian genomes. This diversity is mainly based on two different biological mechanisms: alternative transcription and alternative processing of primary RNA transcripts. Alternative splicing is known to affect more than half of all genes in human, and has been proposed as a primary driver of the evolution of phenotypic complexity in mammals. Recently, we evaluated the prevalence and evolutionary conservation of potential control elements, namely, upstream AUGs and upstream open reading frames, in 5'UTRs of human and mouse genes that are impacted by alternative events. Our findings on selection in alternative and constitutive regions are consistent with the hypothesis that alternative events, related to both processing and transcription initiation, in 5'UTRs of mammalian genes substantially contribute to the regulation of translation. Also we have analyzed the distribution of variable and constitutive nucleotides in different human genome loci producing alternative transcripts. Our results showed that ratios between these two categories of nucleotides are dramatically different for transcript functional domains, such as 5'UTRs, CDSs and 3'UTRs. However, this ratio is stable within the same functional domain in different transcripts. Likely, major variability of 5'UTRs is related to alternative transcription initiation. Although variability of the coding regions, where the ratio of constitutive to alternative nucleotides dramatically higher, is mainly due to alternative

splicing. 3'UTRs are significantly less variable than 5'UTRs and show the same trend of involvement in alternative splicing as CDSs. Negative association between alternative events in 5'UTRs and CDSs could be partially explained by the nature of predominant mechanisms of variability in these functional domains: alternative transcription initiation for 5'UTRs and alternative splicing for CDSs. These two domains demonstrate different patterns of alternative splicing. Interestingly, a higher level of alternative events was found in more ancient genes with phylogenetic hits shared between different phyla of cellular organisms. However, there is no significant difference in phylogenetic patterns between genes with alternative splicing and alternative initiation of transcription. Studies of alternative tissue-specific regulation in mammalian transcriptomes and the role of alternative events in regulation of gene expression level and breadth are in the progress.

### **Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals**

We performed a genome-wide comparative analysis of orthologous sets of mammalian and yeast mRNAs, and revealed distinct patterns of evolutionary conservation at the boundaries of the untranslated regions (UTRs) and coding regions (CDSs). The level of conservation upstream of the start codon and downstream of the stop codon differs from expected and likely reflects hidden functional signals, such as ribosomal filters that could regulate translation by modulating the interaction between mRNAs and ribosomes.

### **Neural networks with miRNA-motivated thermodynamics and composition features improve siRNA design**

653 small interfering RNAs (siRNA) with known efficiencies were analyzed in this study. We compared several parameters for miRNA and siRNA prediction and suggested original set of optimized parameters for siRNA design. A number of common structural features provide evidence of similarity of the miRNA and siRNA mechanisms of action. Based on the constructed set of parameters we used artificial neural networks to predict optimal targets for RNAi experiments (2431 siRNAs published by Novartis). Our neural network model produced better overall correlation between predicted and observed efficiency than other currently employed models trained on the same data set.

### **Genome-scale siRNA prediction**

Using the results of the previous study, we designed and implemented software for predicting optimal siRNA targets for all known human transcripts. We also plan further improvement of this software, in particular, development of the web based interface that would allow to add this software to the standard NCBI services.

### **Analysis of internal loops within the RNA secondary structure in an almost quadratic time**

This work is an improvement of Eppstein's algorithm for evaluating possible internal loops in RNA secondary structure. We proposed a new algorithm that uses Zuker's internal loop penalty function and have run-time of  $O(M \cdot \log^2 L)$ , which improves  $O(M \cdot L)$  time. We also proposed algorithms, each with the run-time of  $O(M \cdot \log^2 L)$ ,

which construct sets of conditionally optimal multi-branch free structures. These algorithms are used in our siRNA design software.

### **A common human COMT haplotype modulates protein expression by altering mRNA secondary structure**

I collaborated with the group from the Center for Neurosensory Disorders (UNC, Chapel Hill) on the study of common genetic variants of catecholamine-O-methyltransferase (COMT). Three common single nucleotide polymorphisms (SNPs) in the coding region of the gene, two of which lead to synonymous changes and one of which leads to a nonsynonymous change, form three major haplotypes. Common genetic variants coding for reduced enzymatic activity, are associated with increased pain sensitivity and the likelihood of developing a persistent musculoskeletal pain condition.

The goal of our study was to understand the mechanism whereby *COMT* haplotypes modulate enzymatic activity. We hypothesized that an interaction between alleles leads to alterations in mRNA secondary structure corresponding to the three haplotypes. RNA folding analyses demonstrated that the major *COMT* haplotypes indeed code for mRNA molecules that substantially differ with respect to their local mRNA secondary structures. The haplotype coding for the lowest enzymatic activity forms an extended local stem-loop structure with the lowest free energy ( $\Delta G$ ). We showed that cells expressing the haplotype coding for this extended stem-loop structure exhibit the lowest amount of enzymatic activity due to deficient protein synthesis, independent from RNA expression and amino acid composition. Furthermore, the original protein and enzymatic activity levels were restored by site directed mutagenesis that destroyed this extended stem-loop structure and converted it into the original stem-loop mRNA structure or by removal of the 5' and 3' untranslated regions which participate in extended stem-loop formation. Collectively, our results provide a new perspective on the functional role and significance of synonymous variants common in the human population.

### **Periodic pattern of mRNA secondary structure imposed by the genetic code and selection in favor of nucleotides G and C in synonymous codons**

Currently I am studying evolution of RNA secondary structures and their influence on the regulation of gene expression. We performed the first transcriptome-scale analysis of the human and mouse mRNA foldings and revealed a universal periodic pattern of nucleotide involvement in mRNA secondary structure in the protein coding regions, which is created by the structure of the genetic code and the unequal use of synonymous codons. Strikingly, the least conserved 3<sup>rd</sup> degenerate codon sites make the biggest contribution to mRNA stability, as compared to the 1<sup>st</sup> and the 2<sup>nd</sup> sites. Furthermore, there is a significant negative correlation between sequence conservation and the free energy of base pairing at the 3<sup>rd</sup> codon sites. Our results convincingly support the hypothesis that redundancies in the genetic code enable transcripts to satisfy requirements for both protein and RNA structure, and suggest that selection in favor of G and C may be operating in synonymous codons to maintain a more stable and ordered mRNA secondary structure.

### **PAST RESEARCH PROJECTS**

The main areas of my research were: structural organization of evolutionary conserved regulatory regions in nucleic acids (in particular, sites of DNA interaction with activators or repressors) (1985-1991), properties of potential sites of RNA-RNA interaction in the course of translation (1992-1995), mutation accumulation experiments in *Drosophila melanogaster* (1995-1997), comparative analysis of different genes and genomes (in particular, of their 5'-untranscribed regions) from different species (1985-1997), genomic sequence comparison (*C.elegans* and *C.briggsae*, *Homo sapiens* and *Mus musculus*), creation of alignment data base and development of new approaches for solving problems of genomic sequence comparison (1999-2006), evolution of RNA structures, RNA folding and selection of stable secondary structures (2004-2007).

### **Comparative analysis of individual genes and genome sequences**

I performed an extensive characterization of *Drosophila* heat-shock genes. This included the search for periodicity in the distribution of A,T-rich motifs and constructing alignments of their 5'-regulatory regions. Drastic differences in the mono- and dinucleotide frequencies between various regions were found. I compared two genomes of *Caenorhabditid* nematodes - *C.elegans* and *C.briggsae* - in order to evaluate mutation rate in different functional regions of genome sequences, namely, exons, introns and spacers. The aim of this work was further understanding of what parts of genomes are neutral and what functional regions are controlled by selection. Similar approach was applied to human - mouse genomic comparison. 100 pairs of complete, orthologous intergenic regions from the human and mouse genomes (average length about 12 000 nucleotides) were aligned and analyzed. The alignments alternate between highly similar segments and dissimilar segments, indicating a wide variation of selective constraint. It was found that the average number of selectively constrained nucleotides within a mammalian intergenic region is at least 2000. This is three fold higher than within a nematode intergenic region and at least two fold higher than the number of selectively constrained nucleotides coding for an average protein. Because mammals possess only two- to three fold more proteins than *Caenorhabditis elegans*, the higher complexity of mammals might be primarily due to of the functioning of intergenic regulatory DNA sequences.

### **Analysis of DNA conservative regulatory regions**

#### ***Statistical analysis of conservative regulatory sites***

Using an original data base of DNA regulatory sites, I have analyzed the frequencies of nucleotides and nucleotide substitutions in conserved DNA regions involved in the regulation of gene expression. It was found that deviations from a consensus tend to cluster in adjacent positions. In particular, transversion are usually accompanied by adjacent compensatory substitutions, while transitions more often occur alone. On the basis of these observations a set of rules describing the pattern of nucleotide substitutions in regulatory sites was proposed. These rules can be used to derive consensus sequences for sets of related regulatory sites.

Comparisons between orthologous intergenic regions of related genomes reveal numerous hits, i.e. pairs of relatively short highly similar sequences that evolved slowly, perhaps due to selective constraint. 2638 hits found within 100 pairs of complete,

orthologous intergenic regions of human and murine genomes were analyzed. All common fragments of hits that align well with many other hits were identified and their classification was constructed. This analysis revealed 20 abundant classes each containing 10 or more fragments. Fragments of the same class may perform the same function, e.g. bind a particular protein. Ten of the abundant classes apparently correspond to known functional consensuses, whereas others may represent novel conserved sites. Thus, large-scale comparative analysis of slowly evolving intergenic sequences can provide valuable insights into their function.

### ***Contributions of individual monomers to biopolymer function***

The functioning of various DNA, RNA and protein sites depends on their sequences. Often, a function may be characterized quantitatively, e. g. by a binding constant, an interaction energy, or an output of reaction product. We showed that it is impossible to determine the absolute contributions of monomers (nucleotides or amino acids) into biopolymer functioning even in the case when the most complete data on biopolymer sequences and their functional efficiencies are available. Algorithm for determination of the relative contributions of monomeric units into biopolymer functioning was developed and applied to several sets of related regulatory sites with quantitatively characterized properties.

### **Investigation of RNA-RNA and RNA-DNA interactions**

#### ***RNA-RNA interaction and translation regulation***

We detected stable intermolecular hybrids between 18S or 28S rRNA and mRNAs of some oncogenes. Similar hybrids were observed by other authors for 5S rRNA and 18S rRNA hybridized with mRNA and human rRNAs and tRNAs. In murine and human 18S rRNAs we identified short regions ("clingers") that can hybridize with an unusually large number of mRNAs from these organisms.

We developed a multiple contact model of intermolecular RNA-RNA interactions. This model describes the formation of stable rRNA-mRNA structures as the result of interactions between numerous short contact regions scattered along the nucleotide sequences. Computer analysis revealed sites potentially involved in mRNA-rRNA and tRNA-rRNA interactions in *E. coli*. The predictions of the model are supported by recent experimental data. The results obtained suggest a possible role for complementary sites in regulation of translation and in enhancement of local concentration of mRNAs and tRNAs near the ribosome.

#### ***Hybridization efficiency of oligonucleotide probes***

New approach for identification of conserved regions in multiple sequence alignments that are thermodynamically suitable for targeting by oligonucleotides was suggested and applied to the HIV genome. We developed a scheme for optimal detection of oligonucleotides hybridization targets common to families of aligned RNA sequences that combines conservation and thermodynamic selection criteria. Our approach involves several steps and employs sequential filtering procedures: (i) creation of a consensus sequence of RNA or DNA from aligned sequence variants with specification of the lengths of fragments to be used as oligonucleotides in the analyses; (ii) selection of fragments in

consensus sequence with homology for the aligned multiple RNA sequence variants, greater than a defined threshold; (iii) selection of DNA oligonucleotides that have pairing potential, greater than a defined threshold, with all variants of the aligned RNA sequences; (iv) elimination of DNA oligonucleotides that have potential for non-specific intra- and inter-molecular interactions. We predicted optimal RNA target regions for consensus oligonucleotides for the HIV-1 genome. They can be used for improvement of oligo-probe based HIV detection techniques. Combination of pre-existing and newly software created in our study can be helpful in design of consensus oligonucleotides with consistently high affinity to RNA targets variants in evolutionary related genes.

### **Software for DNA analysis**

From 1985, I was involved in the development of programs for the analysis of DNA and protein primary structures. The package SAMSON, consisting of 16 computer programs based on original algorithms of sequence comparison and genomic analysis was widely used in Russia (1985-1995). Since 1990 I was a head of a small team of 4 programmers which developed computer programs for advanced biopolymer primary structure analysis. I participated in the development of the OWEN program for aligning long collinear regions of genomes (1999-2001). This program is an interactive tool that represents similarity between sequences as a chain of collinear local similarities. OWEN employs several methods for constructing and editing local similarities and for resolving conflicts between them.

### **Experimental research in molecular biology**

I was involved in several experimental projects. I did a molecular biology project on construction and analysis of *E. coli* plasmids, containing genes of Polyoma viruses and studied regulation of several genes of Polyoma viruses. Also I worked on sequencing and analysis of yeast 18S ribosomal RNA (Institute of Biochemistry and Physiology of Microorganisms, Russian Acad. Sci., Pushchino). I also participated in a biomedicine research project. I worked on a test-system for monitoring biological activity of some natural compounds.

In 1997-1999 I was working on translation regulation of the *COX2* gene in *Saccharomyces cerevisiae*. I studied regulatory signals of the 54-nucleotide *COX2* 5'-untranslated leader sequence involved in translation initiation in yeast mitochondria using mutational and revertant analysis. Results of this work suggest that spacing between regulatory signals is very important for regulation and initiation of translation. Most likely, the control region for interaction of mito mRNAs with ribosomes is similar to that of bacteria and does not exceed 10 nucleotides, although we did not find any specific elements or signals similar to Shine-Dalgarno sequences in the *COX2* mRNA. The presence of complementary matches between the *COX2* mRNA sequence and 15S rRNA lead us to suggestion that stable secondary structures of the 5'UTL allow correct positioning of the start codon on the ribosome.

### **Spontaneous mutations in *Drosophila melanogaster***

In 1995-1997 I took part in a project aimed at measuring the genomic deleterious mutation rate in *Drosophila melanogaster*. We used a purely demographic approach in order to relax selection in two panmictic experimental populations, where each fly

contributed exactly two offspring to the next generation. Comparison with two control populations, one cryopreserved and the other kept under relaxed selection with long generation time, showed that the number of surviving offspring per female assayed under benign and harsh conditions declined by 0.2% and 2.0% per generation, respectively. These data indicate that mutational pressure on fitness may be substantial, provided that the fitness is assayed under harsh, competitive conditions.

## **RESEARCH GOALS**

### **Studies of the RNA world**

#### **Compensatory evolution and RNA secondary structure**

The function of protein and RNA molecules depends on complex epistatic interactions between functional sites. Massive genome sequencing allows studying of epistatic fitness interactions at the genome or transcriptome-scale levels. We showed earlier the existence of a pronounced structural constraint within the coding regions of the human and mouse mRNAs created by the structure of the genetic code and unequal use of synonymous codons. This finding supports the idea that selection may be operating on synonymous codon sites to maintain structural features of mRNA. Our new method for fast secondary structure prediction based on a comparison of nucleotide sequences will be used to study compensatory evolution of miRNA genes and chimp-human mRNAs. Covariations occurring in the helices of the conserved RNA structures will be analyzed and several parameters of the evolution of compensatory mutations will be measured. Results of this transcriptome-scale study will allow testing Kimura's model of compensatory fitness interactions, which assumes that mutations occurring in RNA helices are individually deleterious but become neutral in appropriate combinations.

RNA molecules have a variety of important functions in biological systems, many of which depend on the RNA folding into a precise structure. It is well known that protein synthesis requires participation of tRNAs and rRNAs that have highly conserved structures. An important question is how large is the part of transcriptome where selection on the secondary structure is strong enough to maintain highly conserved structures. Without any doubts, comparative sequence analysis of compensatory substitutions in mRNAs and non-coding RNAs will open new area of studies in evolution of RNA secondary structures.

#### **Analysis of pre-mRNA secondary structures on the transcriptome scale**

Periodicity in mRNA secondary structure facilitates formation of intramolecular helices and compact transcript folding which makes the genetic message more resistant to degradation and modification, while alteration of GC and AU base pairings prevents the formation of strong local secondary structures that may impede translation. Thus, selection seems to operate not for the most stable, but for optimally stable and ordered mRNA secondary structure. These results, derived from our large-scale transcriptome analysis, demonstrate that redundancy of the genetic code allows preservation of both protein and RNA structure, and underscore the importance of the 3<sup>rd</sup> codon sites for the maintenance of mRNA folding. How this periodic pattern affects pre-mRNA secondary structure and the splicing process is intriguing question. We also demonstrated the

existence of specific relaxed secondary structures at the boundaries between 5'UTR-CDS and CDS-3'UTR. I am going to extend my analysis to RNA secondary structure features at the exon-intron boundaries.

### **RNA-RNA interactions**

An important problem in molecular biology is RNA-RNA interactions in the course of splicing of eukaryotic mRNAs. We developed new method for determination of regions of potential intermolecular interactions in different RNA molecules. The next step is the development of the multiple contact model of RNA-RNA intermolecular interaction in processes of translation. The following work is being planned:

New algorithms and software for predicting potential sites of intermolecular interaction based on calculation of probability profiles of intermolecular duplexes formation between mRNA and rRNA oligonucleotide fragments will be developed. The data on prokaryotic and eukaryotic rRNA from different species will be compared and analyzed, taking into consideration known functional regions and predicted secondary structures of these molecules.

The computer predictions will be tested in experiments with DNA oligonucleotides complementary to 18S rRNA (in collaboration with the University of Utah, SLC). The accessibility of 18S rRNA sites in 40S ribosome subunit for oligonucleotide binding, competition between mRNAs and oligonucleotides for interaction with 40S ribosome subunit, and inhibition of protein synthesis by oligonucleotides *in vitro* will be investigated. The 18S rRNA fragments protected from RNAase H digestion will be identified using single-stranded full size DNA analogues of mRNAs.

Three interrelated investigation lines may be followed in this work:

- 1) Development of theoretical models describing binding of rRNA sites to mRNA in the framework of the statistical thermodynamics;
- 2) Development of algorithms and software based on these models, allowing computer identification of potential binding sites on rRNA and mRNA;
- 3) Experimental testing of the computer predictions *in vitro* followed by improvement of models and perfection of algorithms.

The distributions of potential interaction sites along 16S and 18S rRNAs in prokaryotes and eukaryotes will be compared. We also plan further improvement of our new approach to siRNA design. This method is based on the constructed set of parameters and artificial neural networks to predict optimal targets for RNAi experiments and may be applied to all known mammalian transcripts.

### **Evolution of microRNA genes**

It was shown that a subset of mammalian microRNA genes originated from LINE-2 transposable elements and other genome repeats. These repeat-derived microRNAs arise from conventional precursor hairpins. Since the insertion of transposable elements into new genomic sites appears to be one of the driving forces that create new microRNAs during mammalian evolution, it will be interesting to analyze complementarity profiles and other characteristic features of microRNA molecules for selected groups of repeats and transposable elements and to address their potential role in microRNA evolution. Another promising problem is a potential function of some repeats in prokaryotic organisms as a source of noncoding RNAs.

## **Comparative analysis of mammalian genomes:**

### **The function of non-coding DNA sequences**

For decades, researchers focused most of their attention on protein coding genes and proteins. With the completion of the human and mouse genomes and accumulation of data on the mammalian transcriptome, the focus shifts to non-coding DNA sequences, RNA-encoding genes and their transcripts. The historic sequencing and annotation of the complete human genome revealed the complex landscape of mammalian non-coding DNA. Subsequent sequencing of complete genomes of the mouse, rat, dog, cat and cow not only provided genetic platforms for biomedical studies on these model mammalian organisms, but also promoted better understanding of the human genome through comparative analysis. Large-scale sequencing and primary analysis of the mouse and human cDNA libraries provided the first in-depth look into the mammalian transcriptome. These accomplishments allow us to address some yet unanswered questions using genome-wide comparisons. How many genes (separately regulated transcriptional units, encoding distinct transcripts) are there in the mammalian genome, and what is the proportion of the protein coding and non-coding genes? What is the function of non-coding RNA transcripts and non-coding DNA regions? What structural elements in genomes of mammals are responsible for the increased complexity of mammalian organisms?

Comparison of complete genomes of different organisms is obviously very interesting. I am going to conduct a detailed multiple comparison of genomic sequences of mammals (human, mouse, dog, cow, rat and chimp). An important problem is accurate estimation of the number of noncoding RNAs (ncRNAs) in mammalian genomes. In contrast to the reliable and almost complete annotation of protein coding genes in the human genome, comparable information is lacking for ncRNAs. Our new algorithm for RNA secondary structure prediction will allow us to predict the ncRNA potential for mammalian genomes. Previous data on prediction of structured RNA elements are dubious, because conservation of these structures among the mammalian genomes is very low (1:30). It is known that secondary structures of RNA molecules are highly conserved between species and are under selection in favor of a more stable structure than random sequences. I believe that the analysis of widespread conservation of secondary structure, which points to a large number of functional ncRNAs and cis-acting mRNA structures in the human genome, is crucial for understanding of genome composition and evolution in mammals and also important for studies of regulation of transcription and translation.

### **Alternative transcription and alternative splicing in regulation of expression in mammals**

Alternative transcription and alternative splicing generate an enormously increase functional and proteomic diversity in mammals. These two processes have crucial influence on macromolecular and cellular complexity of higher eukaryotes. There is some evidence that majority of alternative processing and transcription events vary between tissues, whereas variation between individuals is less common. The relationship between alternative events and gene expression profile is crucial for understanding of the regulation

of many biological processes. Estimations of the proportion of genes with alternative isoforms in humans differ according to the methodology used, but they could reach ~75% or more, suggesting that frequent alternative events in human genes is the rule. Other mammals exhibit a similar proportion of genes with splicing variants, but alternative isoforms are often lineage-specific, suggesting that alternative events are dynamic processes across evolution. We are trying to understand how alternative events in transcription and processing originated in the course of evolution, through the analysis of different types of mutations in DNA sequences. We are also going to focus on the evaluation of the evolutionary trajectories of alternatively transcribed and processed genes, where we are planning to compare selective pressure on alternative versus constitutive nucleotides in different mammalian species. Ultimately, we are interested in the evolutionary mechanisms allowing new alternative variants to spread and be maintained in populations.

### **Regulation of brain-specific gene expression and function**

Another important problem is prediction and analysis of conserved regulatory signals in the promoter regions of mammalian genes and non-translated regions of mammalian transcripts. Coordinate regulation of transcription and translation may exist for some genes that possess long and complex UTRs, and are involved in signal transduction, cell signaling, morphogenesis, and cell growth control. To reveal these regulatory elements, I am going to conduct multi-species genome-scale analysis of conserved functional sites in intergenic regulatory DNA regions and UTRs. I am also going to perform similar analysis for mammalian genomic sequences with different level of expression.

Genes, specifically involved in the function of the nervous system, sensory perception, and cognition, are of special interest. This area is extremely interesting and also poorly studied. To approach this problem, we selected a superfamily of human protein kinase genes for a pilot study. Using EST analysis, we identified a group of protein kinase genes whose expression is restricted to the nervous system, and identified a number of brain-specific evolutionarily conserved motifs in the 5'UTRs and untranslated regions of these genes that are potentially involved in regulation of brain-specific gene expression (unpublished results). At the next stage, we are going to identify human brain-specific genes at the genome scale, and to extend this analysis to other brain-specific genes, in particular genes encoding receptors of the nervous system, and enzymes involved in metabolism of mediators.

A variety of algorithms will be used in this work. In addition to the standard algorithms for finding regulatory regions, and our algorithms described above, I will also use an original algorithm that searches for regions of drastic changes in DNA sequences, which has been developed with my participation for DNA - protein interaction studies. Preliminary data showed that this algorithm possesses significant advantages over commercially available software for DNA analysis.

### **Comparison of metabolic pathways**

An interesting and challenging problem is comparison of metabolic pathways of different species, inferring these pathways from their genome sequences. This will advance understanding of the evolution of metabolism, the relationship between the evolution of

genome sequence and its function, and the early phylogeny of life. More specifically, I am going to do the following: 1) to reconstruct, to the extent possible, the metabolic pathways for some evolutionary divergent species with known complete genomes, using the data on the functions of proteins; 2) to compare the reconstructed pathways for different organisms and relate their differences to the large-scale differences between their genomic sequences; 3) to derive the possible trajectories of evolutionary transitions between the different metabolic pathways; 4) to determine the importance of horizontal transfer in the evolution of the metabolic pathways; 5) to propose hypotheses on the metabolism in the common ancestor of Eubacteria, Archebacteria, and Eukaryota.